ABSTRACT OF THE DISCLOSURE

Techniques are provided for evenly distributing data items of a particular set of data to a plurality of buckets. The buckets of data items may then be assigned to processes to perform operations on the data items in parallel with the other processes. In one embodiment, the set of data (which may come from tables or be the result set of a previous operation) is divided into a plurality of subsets. For each subset of the plurality of subsets, a sample of data items is randomly selected. The sampling itself may be performed in parallel, with each sampling process using a different seed to randomize its selection of samples. The sampled data items are sorted and ranges are determined based on distribution keys of the sampled data items. The ranges are assigned to buckets, and the data items are then distributed to the buckets assigned to the range into which their distribution key falls.